Docket No.: *324-163* **PATENT**

IN THE UNITED STATES PATENT AND TRADEMARK OFFICE

In re Application of	:	
Johannes HEINECKE	:	Confirmation No.: 5874
U.S. Patent Application No. 10/732,809	:	Group Art Unit: 2644
Filed: December 11, 2003	:	

For: TEXT LANGUAGE IDENTIFICATION

TRANSMITTAL OF CERTIFIED PRIORITY DOCUMENT

Commissioner for Patents
P.O. Box 1450
Alexandria, VA 22313-1450

Sir:

At the time the above application was filed, priority was claimed based on the following application(s):

FRANCE Application No. 02-16250, filed December 17, 2002.

A copy of the priority application is enclosed.

Respectfully submitted,

LOWE HAUPTMAN GILMAN & BERNER, LLP

Allan M. Lowe
Registration No. 19,641

1700 Diagonal Road, Suite 300
Alexandria, Virginia 22314
(703) 684-1111
(703) 518-5499 Facsimile
Date: March 25, 2004
AML/TLH

R E P U B L I Q U E F R A N Ç A I S E

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION

CERTIFICAT D'UTILITÉ - CERTIFICAT D'ADDITION

COPIE OFFICIELLE

Le Directeur général de l'Institut national de la propriété industrielle certifie que le document ci-annexé est la copie certifiée conforme d'une demande de titre de propriété industrielle déposée à l'Institut.

Fait à Paris, le 11 DEC. 2003

Pour le Directeur général de l'Institut
national de la propriété industrielle
Le Chef du Département des brevets

Martine PLANCHE

INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE
SIEGE
26 bis, rue de Saint Petersbourg
75800 PARIS cedex 08
Téléphone : 33 (0)1 53 04 53 04
Télécopie : 33 (0)1 53 04 45 23
www.inpi.fr

ETABLISSEMENT PUBLIC NATIONAL CREE PAR LA LOI N° 51-444 DU 19 AVRIL 1951

1er dépôt

INPI
INSTITUT NATIONAL DE LA PROPRIÉTÉ INDUSTRIELLE
26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 33 (1) 53 04 53 04 Télécopie : 33 (1) 42 94 86 54

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11354*03

REQUÊTE EN DÉLIVRANCE page 1/2

BR1

Cet imprimé est à remplir lisiblement à l'encre noire DB 540 W / 210502

Réservé à l'INPI

REMISE DES PIÈCES
DATE 17 DEC. 2002
LIEU 99
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0216250
DATE DE DÉPÔT ATTRIBUÉE PAR L'INPI 17 DEC. 2002

1 NOM ET ADRESSE DU DEMANDEUR OU DU MANDATAIRE À QUI LA CORRESPONDANCE DOIT ÊTRE ADRESSÉE

CABINET MARTINET & LAPOUX
Conseils en Propriété Industrielle
43 boulevard Vauban
B.P. 405 GUYANCOURT
78055 ST QUENTIN YVELINES CEDEX

Vos références pour ce dossier *(facultatif)* MD/CNET04443

Confirmation d'un dépôt par télécopie	☐ N° attribué par l'INPI à la télécopie
2 NATURE DE LA DEMANDE	**Cochez l'une des 4 cases suivantes**
Demande de brevet	☒
Demande de certificat d'utilité	☐
Demande divisionnaire	☐
Demande de brevet initiale	N° Date
ou demande de certificat d'utilité initiale	N° Date
Transformation d'une demande de brevet européen *Demande de brevet initiale*	☐ N° Date

3 TITRE DE L'INVENTION (200 caractères ou espaces maximum)

Identification de langue d'un texte

4 DÉCLARATION DE PRIORITÉ OU REQUÊTE DU BÉNÉFICE DE LA DATE DE DÉPÔT D'UNE DEMANDE ANTÉRIEURE FRANÇAISE	Pays ou organisation Date N°
	Pays ou organisation Date N°
	Pays ou organisation Date N°
	☐ **S'il y a d'autres priorités, cochez la case et utilisez l'imprimé «Suite»**
5 DEMANDEUR (Cochez l'une des 2 cases)	☒ **Personne morale** ☐ **Personne physique**
Nom ou dénomination sociale	FRANCE TELECOM
Prénoms	
Forme juridique	Société Anonyme
N° SIREN	3 8 0 1 2 9 8 6 6
Code APE-NAF	
Domicile ou siège — Rue	6, Place d'Alleray
Domicile ou siège — Code postal et ville	7 5 0 1 5 PARIS
Domicile ou siège — Pays	FRANCE
Nationalité	Française
N° de téléphone *(facultatif)*	N° de télécopie *(facultatif)*
Adresse électronique *(facultatif)*	
	☐ **S'il y a plus d'un demandeur, cochez la case et utilisez l'imprimé «Suite»**

Remplir impérativement la 2ème page

1er dépôt

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

REQUÊTE EN DÉLIVRANCE
page 2/2

BR2

Réservé à l'INPI

REMISE DES PIÈCES DATE 17 DEC. 2002
LIEU 99
N° D'ENREGISTREMENT NATIONAL ATTRIBUÉ PAR L'INPI 0216250

GB 540 W / 210502

6 MANDATAIRE *(s'il y a lieu)*		
Nom		LAPOUX
Prénom		Roland
Cabinet ou Société		CABINET MARTINET & LAPOUX
N °de pouvoir permanent et/ou de lien contractuel		
Adresse	Rue	43 boulevard Vauban B.P. 405 GUYANCOURT
	Code postal et ville	7 8 0 5 5 ST QUENTIN YVELINES CEDEX
	Pays	FRANCE
N° de téléphone *(facultatif)*		01 30 64 90 09
N° de télécopie *(facultatif)*		01 30 64 90 02
Adresse électronique *(facultatif)*		martinet@wanadoo.fr
7 INVENTEUR (S)		**Les inventeurs sont nécessairement des personnes physiques**
Les demandeurs et les inventeurs sont les mêmes personnes		☐ Oui ☒ Non : **Dans ce cas remplir le formulaire de Désignation d'inventeur(s)**
8 RAPPORT DE RECHERCHE		**Uniquement pour une demande de brevet (y compris division et transformation)**
Établissement immédiat ou établissement différé		☒ ☐
Paiement échelonné de la redevance *(en deux versements)*		**Uniquement pour les personnes physiques effectuant elles-mêmes leur propre dépôt** ☐ Oui ☐ Non
9 RÉDUCTION DU TAUX DES REDEVANCES		**Uniquement pour les personnes physiques** ☐ Requise pour la première fois pour cette invention *(joindre un avis de non-imposition)* ☐ Obtenue antérieurement à ce dépôt pour cette invention *(joindre une copie de la décision d'admission à l'assistance gratuite ou indiquer sa référence)* : **AG**
10 SÉQUENCES DE NUCLEOTIDES ET/OU D'ACIDES AMINÉS		☐ Cochez la case si la description contient une liste de séquences
Le support électronique de données est joint		☐
La déclaration de conformité de la liste de séquences sur support papier avec le support électronique de données est jointe		☐
Si vous avez utilisé l'imprimé «Suite», indiquez le nombre de pages jointes		

11 SIGNATURE DU DEMANDEUR OU DU MANDATAIRE (Nom et qualité du signataire)	**VISA DE LA PRÉFECTURE OU DE L'INPI**
Roland LAPOUX Mandataire CPI/92-1136	C. TRAN

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire.
Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.

Identification de langue d'un texte

La présente invention concerne l'identification de la langue d'un texte qui peut être court et composé de quelques mots, voire d'un seul mot.

L'invention s'applique particulièrement au traitement automatique de langue naturelle au moyen duquel la langue d'un texte écrit est reconnue avant que par exemple le texte soit traduit dans d'autres langues, ou soit synthétisé en un message vocal. Les outils de traitement automatique de langue naturelle, tels que des analyseurs syntaxiques et/ou sémantiques, utilisent des ensembles de données caractérisant une seule langue à la fois, tels qu'un lexique de forme lexicale de base constituant des entrées de dictionnaire ou de lexique, des règles morphologiques et des règles grammaticales, pour une seule langue à la fois. Même si l'outil est capable de traiter n'importe quelle langue, les données sont souvent préparées afin d'analyser une langue à la fois.

L'identification de la langue d'un texte est donc indispensable avant d'analyser linguistiquement le texte.

Selon un autre exemple, l'identification de langue est encore plus nécessaire lorsqu'un texte est écrit dans plusieurs langues par exemple de manière à traduire ce texte multilingue en une langue unique.

Le brevet US 5062143 propose une approche statistique au moyen de trigrammes, c'est-à-dire de suites de trois caractères consécutifs pour identifier la langue utilisée dans un texte. Initialement, pour chaque langue, des trigrammes qui apparaissent le plus fréquemment dans un texte de

cette langue d'une taille raisonnable, par exemple de 500 caractères environ, sont détectés de manière à constituer un ensemble clé de trigrammes. Les trigrammes dont la fréquence d'occurrence est au moins égale à une fréquence prédéterminée constituent l'ensemble clé de cette langue. Par exemple pour un alphabet de 26 lettres et des trigrammes composés de caractères, y compris au moins un espace, l'ensemble clé comprend 80 trigrammes approximativement qui surviennent à une fréquence d'occurrence représentative d'une probabilité assez élevée.

Le texte dont la langue doit être identifiée est ensuite analysé pour le décomposer en trigrammes de manière à y reconnaître et à compter les trigrammes de l'ensemble clé pour une langue donnée. Les trigrammes des ensembles clé pour les autres langues sont également détectés et comptés. La langue pour laquelle le pourcentage de correspondances de trigrammes avec l'ensemble clé respectif est le plus grand et excède une valeur prédéterminée, est réputée être la langue dans laquelle le texte est écrit.

La présente invention n'est pas limitée à des trigrammes, mais utilise des n-grammes typiques et non systématiques. En effet, les trigrammes d'une langue ne sont pas le reflet précis de caractéristiques d'une langue qui permettent de la distinguer par rapport à d'autres langues prédéterminées.

L'identification précédente d'une langue par approche statistique dépend considérablement de la longueur du texte dont la langue est à identifier. Si le texte, telle qu'une phrase, est assez long, l'approche par trigrammes selon le brevet US 5062143 confère un résultat fiable même si le texte contient des mots d'une autre langue. En revanche,

l'identification d'une langue dans une phrase courte au moyen seulement de trigrammes est nettement moins précis d'autant plus que le nombre de langues susceptibles d'être identifiées est grand. Par exemple, la langue de la phrase anglaise "I want to go to Birmingham" peut être identifiée comme polonaise à cause des trigrammes "t-o-espace", et "espace-t-o" qui sont plus fréquents en polonais qu'en anglais.

Au lieu d'identifier la langue d'un document textuel entier, le procédé selon le brevet US 6292772 B1 identifie la langue de mots individuels avec précision. Le procédé d'identification selon ce brevet utilise des n-grammes de caractères d'une longueur quelconque, par exemple des unigrammes, des bigrammes, des trigrammes, et ainsi de suite, et non seulement des trigrammes. Chaque mot est décomposé en un ou plusieurs n-grammes consécutifs de manière à déterminer un premier n-gramme au début du mot, un ou plusieurs n grammes intermédiaires suivants et un dernier n-gramme qui ne se chevauchent pas et qui caractérisent le mot à analyser. Tous ces n-grammes sont comparés à des n-grammes pré-mémorisés d'une langue qui ont été définis statistiquement dans des textes d'apprentissage de la langue.

Ce procédé détermine ainsi la langue à laquelle appartient un mot isolé et est répété pour chacun des mots d'un texte pour identifier la langue de ce texte.

Lorsqu'un mot, c'est-à-dire un modèle de n-grammes, est contenu dans plusieurs langues, des poids respectifs sont attribués à ces langues pour les distinguer. Par exemple si le mot est "de", l'approche statistique sans pondération indique une

probabilité identique pour le français, le néerlandais et l'espagnol, ces trois langues contenant le mot "de". La pondération permet de désigner l'une de ces trois langues alors qu'il n'est pas certain que ce mot dans le contexte d'une phrase appartienne véritablement à cette langue.

L'invention vise à accroître la précision de l'identification d'une langue d'un texte relativement court parmi plusieurs langues d'une manière beaucoup plus fiable et efficace que par les approches statistiques précitées, indépendamment de l'analyse d'un texte d'apprentissage dans chacune des langues à identifier. Elle vise également à réduire la capacité de mémoire dans un dispositif d'identification automatique de langue, comparativement à l'étendue considérable de mémoire requise pour mémoriser tous les trigrammes et plus généralement les n-grammes de chaque langue selon les approches statistiques selon la technique antérieure.

A cette fin, un dispositif identifiant automatiquement la langue d'un texte numérique et ayant mémorisé des chaînes de caractère numériques pour chacune de plusieurs langues prédéterminées, est caractérisé en ce qu'il comprend :

- un moyen pour mémoriser préalablement des premières chaînes de caractère qui sont fréquentes respectivement dans des mots des langues prédéterminées et les caractérisent, et des deuxièmes chaînes de caractère qui sont atypiques respectivement dans les langues prédéterminées,

- un moyen pour analyser des mots extraits du texte numérique afin de constituer pour chaque mot

extrait des chaînes de caractère contenues dans le mot extrait,

- un moyen pour comparer chacune des chaînes de caractère contenues dans le mot extrait aux premières et deuxièmes chaînes de caractère d'une langue déterminée afin qu'à chaque fois qu'une première chaîne de caractère est trouvée dans le mot extrait, un score associé à la langue déterminée augmente d'un premier coefficient dépendant de la position de la première chaîne de caractère trouvée dans le mot extrait, et à chaque fois qu'une deuxième chaîne de caractère est trouvée dans le mot extrait, le score de langue diminue d'un deuxième coefficient respectif associé à la deuxième chaîne de caractère trouvée, et

- un moyen pour comparer les scores du texte associés aux langues prédéterminées (afin de déterminer le plus grand des scores qui identifie la langue du texte.

Toutes les chaînes de caractère contenues dans le mot extrait ont une longueur comprise entre un caractère et le nombre de caractère dans le mot extrait. Typiquement, une première chaîne de caractère est constituée par l'une des chaînes de caractère suivantes dans un mot extrait : un préfixe, un pseudo préfixe, un suffixe, un pseudo suffixe, un infixe, un pseudo infixe

Le premier coefficient d'une première chaîne de caractère dans le mot extrait peut également dépendre de la fréquence de la chaîne de caractère dans la langue déterminée et/ou de la longueur de la chaîne de caractère.

L'identification de langue est encore plus précise lorsque le dispositif comprend un moyen de comparaison pour comparer chacun des mots extraits du

texte avec des mots fréquents dans la langue déterminée et initialement répertoriés dans un moyen de mémoire afin qu'à chaque fois qu'un mot fréquent est trouvé dans le texte, le score de la langue déterminée n'augmente que d'un coefficient dépendant de la longueur du mot fréquent.

D'autres caractéristiques et avantages de la présente invention apparaîtront plus clairement à la lecture de la description suivante de plusieurs réalisations préférées de l'invention en référence aux dessins annexés correspondants dans lesquels :

- la figure 1 est un bloc-diagramme schématique d'un dispositif d'identification automatique de langue selon l'invention ; et
- la figure 2 est un algorithme d'identification automatique de langue selon l'invention.

L'identification automatique de langue selon l'invention est mise en oeuvre sous la forme d'une application spécifique implémentée dans un ordinateur personnel (PC), ou dans un terminal d'usager pouvant être connecté à réseau de téléphonie ou de radiotéléphonie et éventuellement au réseau internet, ou bien dans un serveur qui est interrogeable par des usagers depuis leurs terminaux. L'ordinateur personnel ou le terminal ou le serveur constitue un dispositif d'identification de langue DI selon l'invention dont les principaux composants logiciels relatifs à l'identification automatique de langue selon l'invention sont illustrés à la figure 1.

Un texte qui est initialement saisi dans le ou transmis au dispositif d'identification de langue DI est numérisé en un texte numérique TX formaté par exemple conformément aux spécifications du langage de

documents hypertextes HTML (HyperText Markup Language). Des fichiers d'en-tête et de fin spécifiques sont ajoutés au fichier du texte TX à traiter pour que le fichier de texte soit au format HTML. Le texte TX à traiter est découpé en mots MT par un extracteur 1 qui détecte les espaces entre les mots dans le texte TX. Les mots séparés trouvés dans le texte TX sont mémorisés dans le dispositif DI.

Le dispositif DI comprend essentiellement trois types de base de données 2, 3 et 4 pour mémoriser respectivement des premières chaînes de caractère relativement fréquentes classées dans trois répertoires PRq, SUq et INq, un répertoire de chaînes de caractère atypiques voire improbables ATq, et un répertoire de mots fréquents MFq relatifs à une langue déterminée Lq. Les bases de données 2, 3 et 4 contiennent en pratique chacune 3Q, Q, Q répertoires relatifs respectivement à plusieurs langues L1 à LQ, avec $1 \leq q \leq Q$, où Q est un entier qui peut être au moins égal à 2, typiquement de l'ordre de 10 à 30 environ ou plus. En variante, les bases de données 2, 3 et 4 sont groupées dans un serveur de gestion de base de données (SGBD).

Les trois premiers répertoires PRq, SUq et INq sont relatifs à des morphèmes, syllabes et petites chaînes de caractère CH à 1 à par exemple 6 caractères qui se retrouvent fréquemment dans la langue Lq et qui caractérisent particulièrement la langue Lq par rapport aux autres langues L1 à L(q-1) et L(q+1) à LQ. Le répertoire PRq contient des préfixes, y compris des syllabes et des chaînes de caractère d'attaque non morphologiques, c'est-à-dire des pseudo préfixes, rencontrés fréquemment dans des mots de la langue Lq. Le répertoire SUq contient des

suffixes, y compris des syllabes et des terminaisons non morphologiques, c'est-à-dire des pseudo suffixes, rencontrés fréquemment dans des mots de la langue Lq. Le répertoire INq contient des infixes, y compris des syllabes et des chaînes de caractère non morphologiques, c'est-à-dire des pseudo infixes, rencontrés fréquemment à l'intérieur de mots dans la langue Lq.

Par exemple, pour la langue française, les trois répertoires précédents PRq, SUq et INq contiennent les morphèmes et autres chaînes de caractère caractéristiques suivants :

PRq = de-, re- n'-, l'-, qu'-, d'- ;

SUq = -ais, -aient, -ent, -ée, -ées, -er ;

INq = -ou-, -oi-.

Chaque chaîne de caractère CH inscrite dans l'un des répertoires PRq, SUq et INq dans la base de données 2 est associée à un coefficient respectif CCH qui dépend d'un coefficient PO et optionnellement de deux coefficients FR et LON.

Le coefficient PO dépend de la position de la chaîne de caractère CH dans un mot de la langue Lq, c'est-à-dire si la chaîne de caractère CH constitue un préfixe ou pseudo préfixe, un suffixe ou pseudo suffixe, ou un infixe ou pseudo infixe. Par exemple pour le français, le coefficient PO est le plus grand pour le répertoire de suffixe SUq et égal à 1,4, est intermédiaire pour le répertoire de préfixe PRq et égal à 1,3, et est le plus petit pour le répertoire d'infixe INq et égal à 0,8.

Le coefficient FR dépend de, par exemple est proportionnel à, la fréquence de la chaîne de caractère CH dans la langue Lq. Par exemple, le coefficient FR est la moyenne d'occurrences de la chaîne de caractère CH dans un texte relativement

long de plus de mille mots environ dans la langue déterminée Lq. Par exemple, en français la terminaison "-es" est beaucoup plus fréquente et caractéristique de la langue française que la terminaison "-ions". Par conséquent le coefficient FR pour la terminaison "-es" est plus élevé que le coefficient FR pour la terminaison moins fréquente "-ions". Pour la langue allemande, le coefficient FR est beaucoup plus faible pour chacune des deux terminaisons précédentes.

Le coefficient LON dépend de, par exemple est proportionnel à, la longueur de la chaîne de caractère CH. Par exemple pour un mot de dix caractères, le coefficient LON est égal à 1,3 ou 1,5 lorsque la chaîne de caractère CH contient trois ou cinq caractères.

Le coefficient CCH est typiquement de la forme :

$$CCH = PO(FR + LON).$$

Le quatrième répertoire ATq dans la base de données 3 concerne des caractères et des chaînes de caractère atypiques qui n'appartiennent pas à la langue Lq et qui sont très rarement utilisés dans celle-ci.

Par exemple pour la langue française, le répertoire ATq est composé d'un ensemble de pseudo préfixes atypiques ou improbables, tels que "kr, ge", un ensemble de pseudo infixes atypiques ou improbables tels que "ä, ö, ñ, uu, kk", un ensemble de pseudo suffixes atypiques ou improbables tels que "-ed", et un ensemble de mots atypiques ou improbables.

Chaque chaîne de caractère atypique dans la base de données 3, également désignée par ATq, est associée à un coefficient respectif CAT qui est

d'autant plus élevé que la chaîne de caractère atypique est improbable dans la langue Lq.

Le cinquième répertoire MFq dans la base de données 4 contient les mots les plus fréquents de la langue Lq.

Par exemple le répertoire MFq contient pour la langue française les mots suivants : "de, le, la, les, y, à, il, ou, et".

Chaque mot fréquent également désigné par MFq est associé à un coefficient CMF représentatif de, par exemple proportionnel à, la fréquence dans la langue Lq et éventuellement la longueur du mot fréquent MFq. Par exemple, le coefficient CMF est la moyenne d'occurrences du mot fréquent MFq dans un texte relativement long de plus de mille mots environ dans la langue déterminée Lq.

Le dispositif d'identification de langue DI comprend également un comparateur de mot 5, un analyseur de mot 6, deux comparateurs de chaîne de caractère 7 et 8, un accumulateur de coefficient 9 et un comparateur de score 10. Les modules logiques 5 à 9 sont utilisés pour chacune des langues L1 à LQ. Pour une langue déterminée Lq, les modules logiques 5 à 9 ont les fonctions suivantes.

Le comparateur de mot 5 compare chacun des mots extraits du texte TX par l'extracteur 1 aux mots fréquents MFq contenus dans la base de données 4 relatifs à la langue déterminée Lq. L'analyseur 6 analyse chaque mot extrait MT de manière à constituer toutes les chaînes de caractère CH incluses dans le mot extrait MT et ayant des longueurs comprises entre un caractère et le nombre de caractère du mot extrait. Chacune des chaînes de caractère CH constituées par l'analyseur 6 est appliquée aux

comparateurs 7 et 8 de manière à les comparer respectivement aux premières chaînes de caractère PRq, SUq et INq contenues dans la base de données 2 pour la langue déterminée Lq et aux deuxièmes chaînes de caractère ATq contenues dans la base de données 3 pour la langue déterminée Lq.

Lorsqu'il y a identité des mots ou des chaînes de caractère dans le comparateur 5, ou 7, 8, celui-ci autorise la lecture du coefficient respectif CMF, ou CCH, CAT dans les bases de données 2, 3, 4 via des portes d'autorisation de lecture ET 51, 71, 81 de manière à appliquer le coefficient lu à l'accumulateur de coefficient 9. Les coefficients CMF et CCH sont additionnés à un score Sq relatif à la langue déterminée Lq et mis initialement à zéro, tandis qu'un coefficient de chaîne de caractère atypique CAT est soustrait au score de langue Sq.

Le comparateur de score 10 compare finalement tous les scores de langue S1 à Sq qui ont été déterminés et mémorisés par l'accumulateur 9 afin de déterminer le plus grand score qui identifie la langue du texte TX.

En référence maintenant à la figure 2, préalablement à l'identification automatique de la langue du texte TX saisi dans le ou transmis au dispositif d'identification de langue DI selon l'invention, les cinq répertoires PRq, SUq, INq, ATq et MFq sont constitués dans les bases de données (SGBD) 2, 3 et 4 pour chaque langue déterminée Lq que le dispositif d'identification de langue DI est capable d'identifier, avec $1 \leq q \leq Q$, à une étape initiale E0. Ces répertoires résultent d'une analyse morphologique de la langue Lq.

L'identification de la langue du texte TX composé de mots MT, ou composé simplement d'un seul mot MT, comprend des étapes principales E1 à E14 montrées à la figure 2.

Après avoir mis un indice de langue q à 0, on recherche une identification par rapport à une première langue et plus généralement à une langue déterminée Lq, à la première étape E1. Une variable de score Sq est mise à 0 et l'indice de langue q est augmenté de 1. En fonction des espaces entre les mots, l'extracteur 1 scinde le texte TX à traiter en des mots extraits MT, à l'étape E2.

Pour chaque mot extrait MT du texte TX, les étapes suivantes E3 à E11 sont effectuées.

A l'étape E3, le comparateur de mot 5 interroge le répertoire de mot fréquent MFq de la langue Lq pour y retrouver éventuellement le mot MT. Si le mot MT appartient au répertoire MFq, le comparateur de mot 5 autorise la lecture du coefficient CMF associé au mot MT via la porte 51 afin que l'accumulateur 9 augmente le score Sq du coefficient lu CMF, à l'étape E4. Plus le mot MT est fréquent dans la langue Lq et est éventuellement long, plus le coefficient CMF est élevé. Puis le procédé d'identification passe à l'étape E11.

Si à l'étape E3 le comparateur 5 ne considère pas le mot MT comme un mot fréquent de la langue Lq, le comparateur 5 active l'analyseur de mot 6. L'analyseur 6 analyse le mot extrait MT pour constituer et mémoriser soit des chaînes de caractère CH incluses dans le mot extrait MT telles que définies respectivement dans les répertoires inclus dans les bases de données 2 et 3, soit toutes les chaînes de caractère CH incluses dans le mot extrait MT et ayant des longueurs comprises entre un

caractère et le nombre de caractère du mot extrait, à l'étape E5. Selon la dernière variante, les chaînes de caractère CH contenues dans le mot extrait MT et trouvées dans les répertoires PRq, SUq et INq peuvent se recouvrir partiellement, contrairement aux n-grammes selon l'approche du brevet US 6292772 B1 déjà commenté. Par exemple, lorsque le mot traité MT est "aiment" en français, les chaînes de caractère "ment" et "ent" placées dans le répertoire de pseudo suffixes SUq se recouvrent dans le mot traité. Selon un autre exemple, l'infixe "oi" et le pseudo suffixe "is" du mot traité "vois" se recouvrent.

L'analyseur 6 applique ensuite successivement les chaînes de caractère CH constituées pour le mot extrait MT aux comparateurs de chaîne de caractère 7 et 8.

Le comparateur 7 recherche dans les répertoires PRq, SUq et INq de la base de données 2 les premières chaînes de caractère CH qui sont contenues dans le mot MT, à l'étape E6. Pour chaque chaîne de caractère CH dans le mot extrait MT appartenant à l'un des répertoires PRq, SUq et INq, le comparateur de chaîne de caractère 7 autorise la lecture du coefficient CCH = PO(FR + LON) associé à la chaîne de caractère CH via la porte 71 afin que l'accumulateur 9 augmente le score Sq du coefficient lu CCH, à l'étape E7. Les coefficients PO, FR et LON dépendent respectivement de la position de la chaîne de caractère CH dans le mot extrait MT, de la fréquence de la chaîne de caractère CH dans la langue déterminée Lq, et de la longueur de la chaîne de caractère CH.

Que la chaîne de caractère CH dans le mot extrait MT ait été trouvée ou non dans les répertoires PRq, SUq et INq, et parallèlement aux étapes E6 et E7, le comparateur de chaîne de

caractère 8 recherche dans le répertoire ATq de la base de données 3 les chaînes de caractère CH qui sont contenues dans le mot MT, à l'étape E8. Pour chaque chaîne de caractère CH dans le mot extrait MT appartenant au répertoire ATq, le comparateur 8 autorise la lecture du coefficient CAT associé à la chaîne de caractère CH via la porte 81 afin que l'accumulateur 9 diminue le score Sq du coefficient lu CAT, à l'étape E9.

Comme indiqué à l'étape E10, les étapes E6 à E9 sont répétées pour chacune des chaînes de caractère CH contenues dans le mot extrait MT. Le score Sq pour la langue déterminée Lq est ainsi augmenté de la somme des coefficients CCH et diminué de la somme des coefficients CAT relatives aux chaînes de caractère CH incluses dans le mot extrait MT et respectivement trouvées dans les répertoires des bases de données 2 et 3, ou augmenté du coefficient CMF si le mot extrait MT a été trouvé dans le répertoire MFq de la base de données 4, après l'étape E10.

Comme indiqué à l'étape E11, si le texte traité TX contient d'autres mots, les étapes E2 à E10 sont exécutées pour chacun des mots extraits du texte TX. Quand tout le texte TX a été traité, si le score Sq pour la langue Lq est négatif à cause de l'application de l'étape E9, le score Sq peut être mis à 0 à l'étape E12. Le score Sq pour la langue Lq est mémorisé dans le dispositif DI qui, à l'étape E13, vérifie que le score d'une autre langue doit être déterminé en exécutant les étapes E1 à E12.

Lorsque les scores S1 à SQ de toutes les langues L1 à LQ ont été déterminés pour le texte traité TX, le comparateur de score 10 compare les Q scores afin de déterminer le plus grand score Sup(Sq) qui identifie la langue du texte TX, à l'étape E14. Puis

par exemple un identificateur de la langue du texte ainsi identifiée est transmis par le dispositif DI à un analyseur syntaxique et sémantique de manière à traduire le texte TX en une langue maternelle d'un utilisateur, ou bien à le convertir sous forme orale, ou bien encore à l'utiliser pour indexer le texte en fonction de sa difficulté de compréhension ou générer des trous pour en constituer un exercice à trous pour l'apprentissage de la langue identifiée.

Le cas échéant à l'étape 14, le dispositif DI affiche à l'utilisateur le score de toutes les langues L1 à Lq trié selon l'ordre décroissant. Le nombre de langue Q n'est a priori pas limité.

Si le texte est constitué par un ou plusieurs mots appartenant à deux langues différentes pour lesquelles les coefficients sont respectivement identiques pour les deux langues, les scores pour ces deux langues sont identiques, et non différents selon la technique antérieure.

L'identification de langue selon l'invention est particulièrement performante par rapport à la technique antérieure lorsque le texte est relativement court. Toutefois, en variante, l'identification de langue selon l'invention peut être combinée avec une identification de langue par approche statistique connue, en appliquant l'identification de langue selon l'invention lorsque le texte contient un nombre de caractères ou un nombre de mots inférieur à un seuil prédéterminé, et en appliquant l'identification connue par approche statistique lorsque le nombre de caractères ou de mots dans le texte traité TX est supérieur au seuil prédéterminé.

REVENDICATIONS

1 - Dispositif identifiant automatiquement la langue d'un texte numérique (TX) et ayant mémorisé des chaînes de caractère numériques pour chacune de plusieurs langues prédéterminées (L1-LQ), caractérisé en ce qu'il comprend :

- un moyen (2, 3) pour mémoriser préalablement des premières chaînes de caractère (PRq, SUq, INq) qui sont fréquentes respectivement dans des mots des langues prédéterminées et les caractérisent, et des deuxièmes chaînes de caractère qui sont atypiques (ATq) respectivement dans les langues prédéterminées,

- un moyen (6) pour analyser des mots extraits du texte numérique (TX) afin de constituer (E5) pour chaque mot extrait (MT) des chaînes de caractère (CH) contenues dans le mot extrait,

- un moyen (7, 8, 9) pour comparer chacune (CH) des chaînes de caractère contenues dans le mot extrait (MT) aux premières et deuxièmes chaînes de caractère (PRq, SUq, Inq ; ATq) d'une langue déterminée (Lq) afin qu'à chaque fois qu'une première chaîne de caractère est trouvée (E6) dans le mot extrait, un score (Sq) associé à la langue déterminée (Lq) augmente (E7) d'un premier coefficient (CCH) dépendant de la position de la première chaîne de caractère trouvée dans le mot extrait, et à chaque fois qu'une deuxième chaîne de caractère est trouvée dans le mot extrait, le score de langue (Sq) diminue (E9) d'un deuxième coefficient respectif (CAT) associé à la deuxième chaîne de caractère trouvée, et

- un moyen (10) pour comparer les scores du texte associés aux langues prédéterminées (L1-LQ) afin de déterminer le plus grand des scores qui identifie la langue du texte.

2 - Dispositif conforme à la revendication 1, dans lequel une première chaîne de caractère est constituée par l'une des chaînes de caractère suivantes dans un mot extrait : un préfixe, un pseudo préfixe, un suffixe, un pseudo suffixe, un infixe, un pseudo infixe.

3 - Dispositif conforme à la revendication 1 ou 2, dans lequel le premier coefficient (CCH) d'une première chaîne de caractère (CH) dans le mot extrait (MT) dépend de la fréquence de la chaîne de caractère dans la langue déterminée (Lq).

4 - Dispositif conforme à l'une quelconque des revendications 1 à 3, dans lequel le premier coefficient (CCH) d'une première chaîne de caractère (CH) dans le mot extrait (MT) dépend de la longueur de la chaîne de caractère.

5 - Dispositif conforme à l'une quelconque des revendications 1 à 4, dans lequel le premier coefficient (CCH) d'une première chaîne de caractère (CH) dans le mot extrait (MT) est déterminé selon la relation suivante :

PO (FR + LON),

dans laquelle PO est un coefficient dépendant de la position de la première chaîne de caractère dans le mot extrait, FR est un coefficient dépendant de la fréquence de la première chaîne de caractère dans la langue déterminée (Lq), et LON est un coefficient dépendant de la longueur de la première chaîne de caractère.

6 - Dispositif conforme à l'une quelconque des revendications 1 à 5, caractérisé en ce qu'il comprend un moyen de comparaison (5, 9) pour comparer chacun des mots extraits (MT) du texte (TX) avec des mots (MFq) fréquents dans la langue déterminée (Lq) et initialement répertoriés dans un moyen de mémoire (4) afin qu'à chaque fois qu'un mot fréquent est trouvé dans le texte, le score (Sq) de la langue déterminée (Lq) n'augmente que d'un coefficient (CMF) dépendant de la longueur du mot fréquent.

1/2

FIG. 1

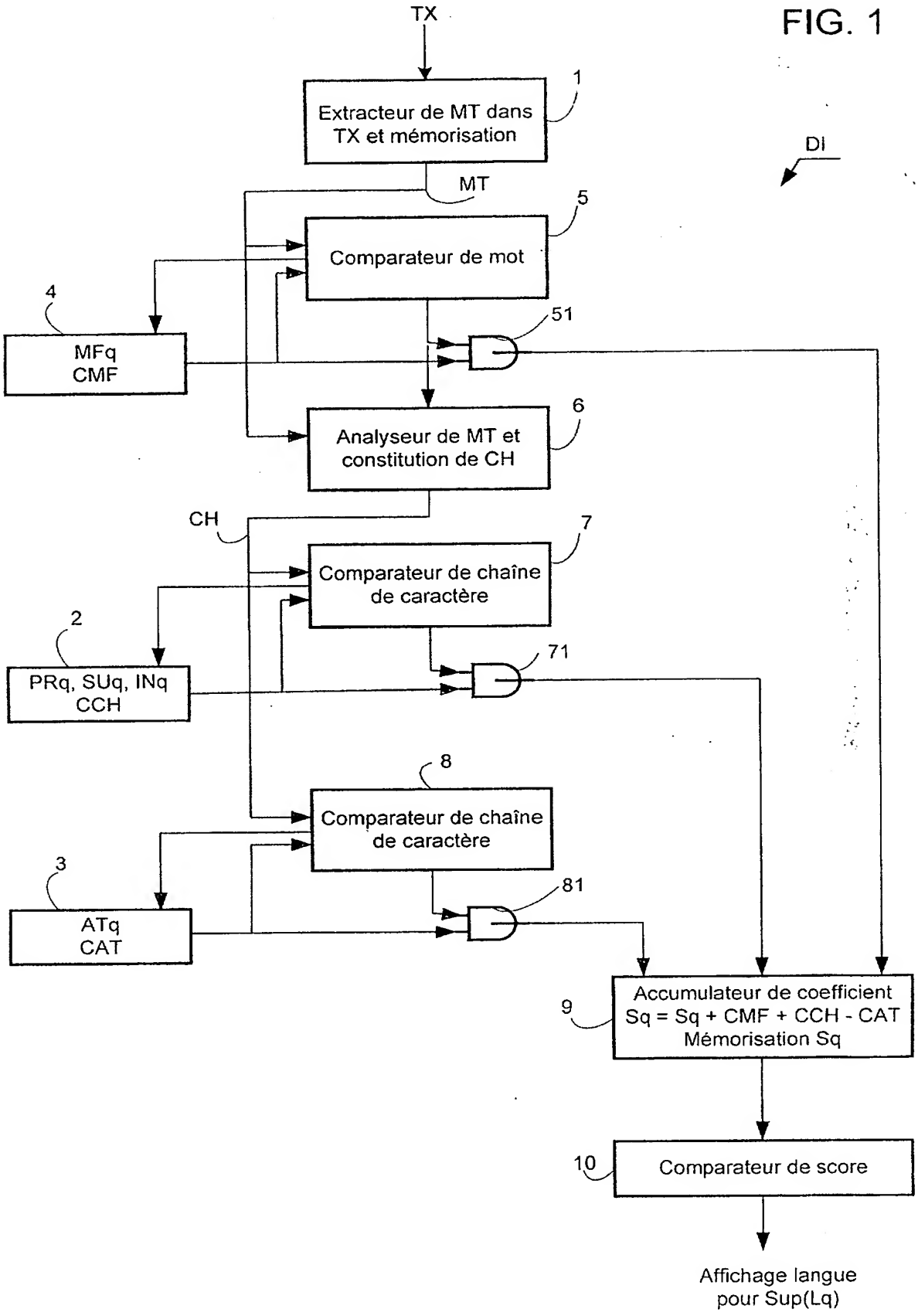
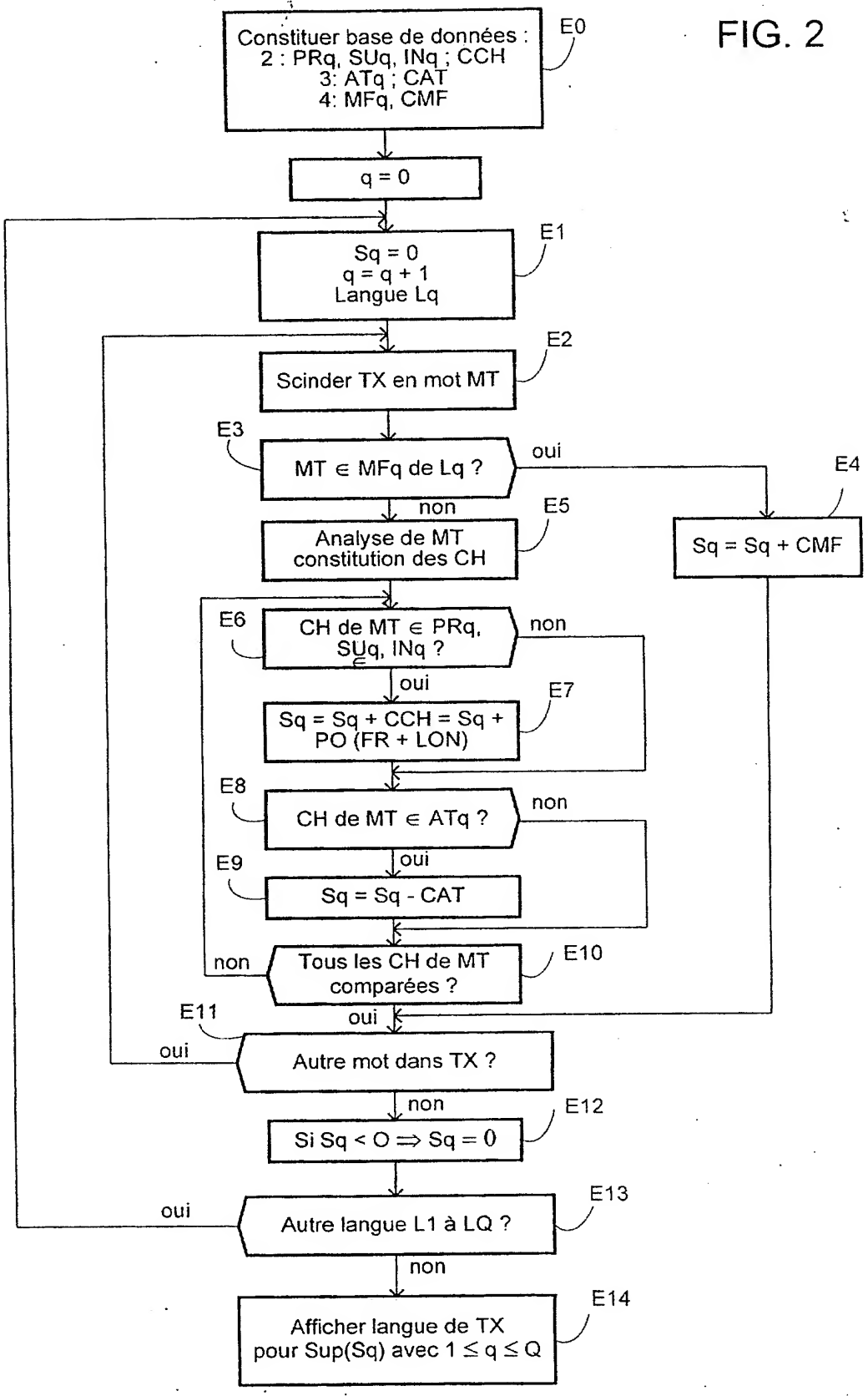

TX
Extracteur de MT dans TX et mémorisation
1
DI
MT
5
Comparateur de mot
4
MFq
CMF
51
6
Analyseur de MT et constitution de CH
CH
7
Comparateur de chaîne de caractère
2
PRq, SUq, INq
CCH
71
8
Comparateur de chaîne de caractère
3
ATq
CAT
81
9
Accumulateur de coefficient
Sq = Sq + CMF + CCH - CAT
Mémorisation Sq
10
Comparateur de score
Affichage langue pour Sup(Lq)

FIG. 2


Constituer base de données :
2 : PRq, SUq, INq ; CCH
3: ATq ; CAT
4: MFq, CMF
E0
q = 0
Sq = 0
q = q + 1
Langue Lq
E1
Scinder TX en mot MT
E2
MT ∈ MFq de Lq ?
E3
oui
E4
non
Analyse de MT
constitution des CH
E5
Sq = Sq + CMF
CH de MT ∈ PRq,
SUq, INq ?
E6
non
oui
Sq = Sq + CCH = Sq +
PO (FR + LON)
E7
CH de MT ∈ ATq ?
E8
non
oui
Sq = Sq - CAT
E9
non
Tous les CH de MT
comparées ?
E10
E11
oui
oui
Autre mot dans TX ?
non
Si Sq < O ⇒ Sq = 0
E12
oui
Autre langue L1 à LQ ?
E13
non
Afficher langue de TX
pour Sup(Sq) avec 1 ≤ q ≤ Q
E14

reçue le 24/01/03

INPI
INSTITUT NATIONAL DE LA PROPRIETE INDUSTRIELLE

DÉPARTEMENT DES BREVETS

26 bis, rue de Saint Pétersbourg
75800 Paris Cedex 08
Téléphone : 33 (1) 53 04 53 04 Télécopie : 33 (1) 42 94 86 54

BREVET D'INVENTION
CERTIFICAT D'UTILITÉ

Code de la propriété intellectuelle - Livre VI

cerfa N° 11235*03

DÉSIGNATION D'INVENTEUR(S) Page N° 1../1..

INV

(À fournir dans le cas où les demandeurs et les inventeurs ne sont pas les mêmes personnes)

Cet imprimé est à remplir lisiblement à l'encre noire

DB 113 @ W / 270601

Vos références pour ce dossier *(facultatif)*	MD/CNET04443
N° D'ENREGISTREMENT NATIONAL	

TITRE DE L'INVENTION (200 caractères ou espaces maximum)

Identification de langue d'un texte

LE(S) DEMANDEUR(S) :

FRANCE TELECOM
6 Place d'Alleray
75015 PARIS
FRANCE

DESIGNE(NT) EN TANT QU'INVENTEUR(S) :

1 Nom		HEINECKE
Prénoms		Johannes
Adresse	Rue	22 rue du Hingar
	Code postal et ville	22300 LANNION
Société d'appartenance *(facultatif)*		
2 Nom		
Prénoms		
Adresse	Rue	
	Code postal et ville	
Société d'appartenance *(facultatif)*		
3 Nom		
Prénoms		
Adresse	Rue	
	Code postal et ville	
Société d'appartenance *(facultatif)*		

S'il y a plus de trois inventeurs, utilisez plusieurs formulaires. Indiquez en haut à droite le N° de la page suivi du nombre de pages.

DATE ET SIGNATURE(S) DU (DES) DEMANDEUR(S) OU DU MANDATAIRE (Nom et qualité du signataire)

Roland LAPOUX
Mandataire
CPI/92-1136

Le 16 Décembre 2002

La loi n°78-17 du 6 janvier 1978 relative à l'informatique, aux fichiers et aux libertés s'applique aux réponses faites à ce formulaire.
Elle garantit un droit d'accès et de rectification pour les données vous concernant auprès de l'INPI.